Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2001 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2001

Data Mining: A Practical Perspective

Joyce Jackson
University of South Carolina-Columbia

DATA MINING: A PRACTICAL PERSPECTIVE

Joyce Jackson
Management Science Department
The Darla Moore School of Business
University of South Carolina – Columbia
joyce.jackson@sc.edu

Abstract

This tutorial provides an overview of the data mining process. Methodological considerations are discussed and illustrated. The tutorial provides a basic understanding of how to plan, evaluate and successfully refine a data mining project particularly in terms of model building and model evaluation.

This tutorial is intended for those with little or no previous experience in data mining. The approach is practical and conceptually sound in order to appeal to both practitioners and academicians.

Introduction

Databases today can range in size into the terabytes. Despite the development of technology to support huge databases, the rapid spread of computerization in all industries presents users with the problem of interpreting vast amounts of data. Within these masses of data lies hidden information that could be of strategic importance. Three tools that have emerged in recent years to address these phenomena are Data Mining, Data Warehousing and OLAP (On-Line Analytical Processing).

Data Mining

The objective of data mining is to identify valid novel, potentially useful, and understandable correlations and patterns in existing data (Chung and Gray 1999). Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines including numerical analysis, pattern matching and areas of artificial intelligence such as machine learning, neural networks and genetic algorithms. While many data mining tasks follow a traditional, hypothesis driven data analysis approach, it is commonplace to employ an opportunistic, data driven approach that encourages the pattern detection algorithms to find useful trends, patterns, and relationships. Regardless of which approach is used, traditional or opportunistic, it must be remembered that data mining is very much a data-oriented process without a strong theoretical background. Therefore, it's procedures and subsequent analyses should be approached with a balance of zeal and vigilance.

Data Mining and Data Warehousing

The data mining database could be a logical or a physical subset of a data warehouse. However, a data warehouse is not a requirement for data mining. Building a large data warehouse that consolidates data from multiple sources, resolves data integrity problems, and loads the data into a database, can be an enormous task, sometimes taking years and costing millions of dollars (Gray and Watson, 1998). If a data warehouse is not available, the data to be mined can be extracted from one or more operational or transactional databases, or data marts.

Data Mining and OLAP

Data mining and OLAP are different tools than can be used to complement one another. OLAP is part of the spectrum of decision support tools. OLAP databases are organized along the different *dimensions* of a business such as time, product type, and

geography. This multidimensional structure is often called a *cube* (even when there are more than three dimensions). Multidimensional systems are powerful resources for reporting and investigating data in a deductive manner. While OLAP pre-summarizes data along specific dimensions, data mining thrives on detail. Again, data mining and OLAP can be used to complement one another. For example, prior to acting on the pattern uncovered by data mining, an analyst may need to determine the implications of using the discovered pattern in governing a decision.

What Data Mining Can Do

While the term data mining is often used rather loosely, it is generally a term that's used for a specific set of activities, all of which involve extracting meaningful new information from data. The activities are: classification, estimation, prediction, clustering, affinity grouping, and description. The first three activities-classification, estimation and prediction are examples of *directed* data mining. In directed data mining, the goal is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. The last three activities – clustering, affinity grouping, and description are examples of *undirected* data mining. In undirected data mining, no variable is singled out as the target variable; the goal is to establish a relationship among all the variables.

What Data Mining Cannot Do

Data mining is a tool; it does not eliminate the need to know the business, to understand the data, or to understand the analytical methods involved. It must be remembered that the predictive relationships found via data mining are not necessarily *causes* of an action or a behavior. Causal inference from uncontrolled convenience samples, such as those used in data mining, are subject to several sources of error such as latent variables, sample selection bias and model equivalence (Glymour and Madigan 1996) . Further, data mining assists analysts with finding patterns and relationships in the data – it does not indicate the value of the patterns to the organization. The patterns uncovered by data mining must be verified and validated in an appropriate context.

Tutorial Outline

The following is an outline of the tutorial, followed by a brief discussion of each major section.

I) The Business Imperative
- i) DM as a Research Tool
- ii) DM for Process Improvement
- iii) DM for Marketing
- iv) DM for Customer Relationship Management

II) The Technical Imperative
- i) DM and Machine Learning
- ii) DM and Statistics
- iii) DM and Decision Support
- iv) DM and Information Technology

III) Methodological Considerations
- i) SAS - The SEMMA Process
- ii) SPSS - The 5 A's Process
- iii) CRISP-DM – the *de facto* standard for industry

IV) Illustration of a Data Mining Process Methodology
- i) Business Understanding
- ii) Data Understanding
- iii) Data Preparation
- iv) Modeling
- v) Evaluation
- vi) Deployment

Tutorial Overview

The Business Imperative

Data mining can be used as a vehicle to increase profits by reducing costs and/or raising revenue. This section elaborates on a few of the common ways in which data mining can accomplish those objectives such as: lowering costs at the beginning of the product life cycle during research and development; determining the proper bounds for statistical process control methods in automated manufacturing processes; eliminating expensive mailings to customers who are unlikely to respond to an offer during a marketing campaign; facilitating one-to-one marketing and mass customization opportunities in customer relationship management.

The Technical Imperative

Data mining uses the classical statistical procedures such as logistic regression, discriminant analysis and cluster analysis, as well as the newer techniques such as neural networks, decision trees and genetic algorithms. An overview of the applicability of these procedures and techniques will be discussed. Next, a distinction will be drawn between data mining and Decision Support Systems. Finally, the advances in server hardware architecture will briefly be addressed. Important developments such as RISC (Reduced Instruction Set Processing), SMP (Symmetric Multi Processing) and MPP (Massively Parallel Processors) provide data mining with the large amounts of fast disk storage and significant processing power that's needed.

Methodological Considerations

Many data mining process methodologies are available. However, the various steps do not differ much from methodology to methodology. Two popular methodologies used by two popular data mining tools are the SEMMA process for SAS Enterprise Miner and the 5 A's process for SPSS Clementine. CRISP-DM, however, has evolved to become the *de facto* industry standard. CRISP-DM was conceived in late 1996 and is non-proprietary, documented and freely available. It was developed using input from more than 200 data mining users and data mining tool and server providers and is designed to provide a generic process model that can be specialized according to the needs of any particular company or industry.

An Illustration of a Data Mining Process Methodology

The CRISP-DM Data Mining Process Methodology will be illustrated. CRISP-DM was designed to provide guidance to data miner "beginners", however, the industry's initial use of the methodology confirmed that it is a valuable aid to beginners and advanced data miners alike (Colin 2000). The primary focus of the illustration will be on the heart of the data mining process methodology: building and evaluating the model, which includes, but is not limited to:

- Creating the model set
- Developing the model
- Training the model
- Validating the model
- Scoring the model
- Evaluating the performance of the model

References

Berry, M. J., Linoff, G. S. *Mastering Data Mining: The Art and Science of Customer Relationship Management*. Wiley Computer Publishing, New York, 2000.

Chung, H. M., Gray, P. "Special Section: Data Mining". *Journal of Management Information Systems,* (16:1), 1999, pp. 11-17.

Colin, S., "The CRISP-DM Model: The New Blueprint for Data Mining", Journal of Data Warehousing, (5:4), Fall 2000, pp. 13-22.

Gray, P., Watson, H.J., *Decision Support in the Data Warehouse,* Upper Saddle River, N.J. (1998).

Glymour, C., Madigan D., et al, "Statistical Inference and Data Mining". *Communications of the ACM,* (39:11), 1996, pp. 35-41.

Wells, M. T., "Feature Extraction Construction and Selection: A Data Mining Perspective". *Journal of the American Statistical Association,* (94:448), 1999, p. 1390.